AD (1)

SOME NUMERICAL EXPERIMENTS USING NEWTON'S METHOD FOR NONLINEAR PARABOLIC AND ELLIPTIC BOUNDARY-VALUE PROBLEMS

Richard Bellman
Mathematics Department
Mario Juncosa
Computer Sciences Department
Robert Kalaba
Electronics Department
The RAND Corporation

P-2200

19

SOME NUMERICAL EXPERIMENTS USING NEWTON'S METHOD FOR NONLINEAR PARABOLIC AND ELLIPTIC BOUNDARY-VALUE PROBLEMS

Richard Bellman
Mathematics Department
Mario Juncosa
Computer Sciences Department
Robert Kalaba
Electronics Department
The RAND Corporation

P-2200

23 January 1961

Reproduced by

The RAND Corporation • Santa Monica • California

The views expressed in this paper are not necessarily those of the Corporation

SUMMARY

Using a generalization of Newton's method, a nonlinear parabolic equation of the form $u_t - u_{xx} = g(u)$, and a nonlinear elliptic equation $u_{xx} + u_{yy} = e^u$, are solved numerically. Comparison of these results with results obtained using the Picard iteration procedure show that in many cases the quasilinearization method offers substantial advantages in both time and accuracy.

u sub t - u sub xx = g(u),

u sub xx + u sub yy = e superscript u,

SOME NUMERICAL EXPERIMENTS USING NEWTON'S METHOD FOR NONLINEAR PARABOLIC AND ELLIPTIC BOUNDARY-VALUE PROBLEMS

Richard Bellman
Mario Juncosa
Robert Kalaba

1. INTRODUCTION

The numerical treatment of initial-value problems in ordinary differential equations on an electronic digital computer usually is no more involved in the nonlinear case than in the linear one. In the handling of boundary-value problems this is not so. In the linear case, when a solution exists, the applicability of the superposition principle provides a decided advantage in boundary-value problems in that it leads to solving at most a number of initial-value problems equal to the order of the system. On the other hand, in the nonlinear case, one of the possibilities for solution is to resort to an iterative technique which replaces the problem with a sequence of linear problems in which one can use the superposition principle.

In solving the differential equation

$$(1) \qquad L(u) = f(u)$$

where L is a linear ordinary differential operator of at least second order and conditions are prescribed to be satisfied by $u(x)$ at at least two points, one may linearize by using Picard's method which introduces a sequence of functions $\{u^{(k)}(x)\}$ which satisfy the same boundary conditions as $u(x)$ and the linear ordinary inhomogeneous differential equation

(2) $$Lu^{(k+1)} = f(u^{(k)}).$$

When the sequence $\{u^{(k)}(x)\}$ converges the convergence is linear, i.e.

(3) $$u^{(k+1)} - u = O(u^{(k)} - u)$$

as $k \rightarrow \infty$.

However, if $f(u)$ is differentiable, we can linearize in a different way. If one replaces the right hand side of (2) by

$$f(u^{(k)}) + (u^{(k+1)} - u^{(k)})f'(u^{(k)})$$

then we have again a linear inhomogeneous ordinary differential equation

(4) $$Lu- f'(u^{(k)})u^{(k+1)} = f(u^{(k)}) - u^{(k)}f'(u^{(k)})$$

which results in a sequence $\{u^{(k)}(x)\}$ which, when convergent, is usually quadratically convergent, i.e.

(5) $$u^{(k+1)} - u = O((u^{(k)} - u)^2)$$

as $k \rightarrow \infty$.

The idea of the use of second-order convergent iterative procedures for solving systems of equations other than algebraic or transcendental is not new. However, except for the work of Hestenes [1] and Stein [2] and a brief mention in Milne's book, sec. 49 [3], it seems to have received only scant attention in the American literature on numerical

analysis compared to that in foreign publications. In 1905, only a few decades after Picard's work, Chaplygin [4] presented what amounts to Newton's method for approximate integration of differential equations. More recently general functional-analytic treatments of Newton's method and some of its variants in a Banach space have been given by a number of authors, notably Kantorovich [5,6], Zagadskii, Mysovskikh, Fenyö, Collatz, Schröder [7], Bartle, and Stein.
References to most of these and to others can be found in [2,6,7].

From the practical point of view, in spite of all this analytic treatment, because of the effects of truncation errors, round-off errors, crude bounds on the higher order derivatives, the labor of computing higher order derivatives or differences for the nigher order methods, the accuracy desired in the final approximation to the solution, and the word length and computing mode of the machine available for the computation the efficiency of the higher order convergent methods vs. that of lower order methods cannot be decided solely on the basis of the order of convergence. An analysis of the interaction of such effects is generally far more difficult than that of simply determining orders of convergence. This difficulty, if not impossibility, justifies some numerical experimentation. Although Kantorovich, Collatz, and Schröder, give some examples of applications to eigenvalue problems, integral equations, and differential equations (mainly ordinary) and Hestenes and Stein give some applications to the calculus of variations, the body of experimental results in the American literature is still quite small.

For ordinary differential equations there is quite a large number of stable methods of numerical integration whose truncation errors are of fairly high order in the integration step size. Consequently, by the choice of one of such methods the interaction between the truncation errors and the rate of convergence of the numerical solutions of either (2) or of (4) as a function of k can be kept quite small. Hence the computational effort to obtain numerical approximations to the solution of (1) is largely determined by whether one chooses to solve the linear differential equations (2) or (4), thus comparing quadratic convergence (5) and some extra computation in the evaluation of $f'(u^{(k)})$ in (4) with linear convergence (3) and no evaluation of derivatives of f in (2). This, indeed, has been compared by one of the authors for a simple nonlinear second order ordinary differential equation with two point boundary conditions. See [8] wherein is given a novel relatively general derivation of (4) utilizing the operation of function maximization. See also [12] for more general applications.

For elliptic and parabolic partial differential equations, however, the stable numerical methods commonly used for solution usually have truncation errors of low order in the mesh sizes. Furthermore, the numerical treatment of elliptic and multi-space-dimensional parabolic cases almost always result in systems of algebraic equations which are the linearized discrete analogs of the partial differential equations and are usually solved by an iterative procedure such as relaxation or overrelaxation. Consequently, regardless of whether one chooses (2) or (4) for the linearization of (1), there is considerable interaction between the truncation errors of the discrete analog, the rate of convergence (3) or (5), depending on the choice of either (2) or (4), and the rate of convergence of the relaxation or overrelaxation procedure.

As a result an analytic representation of realistic bounds on the total error is very difficult, if not impossible, to achieve. The purpose of this note is to present the results of some experiments using (2) and (4) on two cases each of a parabolic partial differential equation and an elliptic one. These cases in each type differing only in the boundary conditions clearly show how markedly the superiority of the method (4) over (2) is affected by the above-mentioned interaction as the boundary conditions are changed.

2. THE PARABOLIC CASE

In the interest of simplicity the experiments on the parabolic case were carried out on the numerical solution of

$$(6) \qquad Lu = u_t - u_{xx} = (1 + u^2)(1 - 2u)$$

over two different triangles: $0 \leq t \leq 1 - x$, $0 \leq x \leq 1$, and $0 \leq t \leq 1.5 - x$, $0 \leq x \leq 1.5$. The boundary conditions in each case were so chosen as to give the solution

$$(7)\qquad u(x,t) = \tan(x + t).$$

That this solution is unique can be established easily by classical techniques.

Solutions of the differential equation analogs of the Picard iteration (2) and of the Newton procedure (4) for the differential equation (6) were compared. The numerical solutions in each case were obtained over the points $(m\Delta t, n\Delta t)$ of a grid superimposed on the respective triangles. In order to avoid the possibility of numerical instability (since we used $\Delta x = \Delta t = .01$ in the experiments) we used the Crank-Nicolson difference operator [9,10] for the discrete analog of (6). Thus, if $u_{m,n}$ designates the solution of a difference equation analog of (6) at the lattice point $(m\Delta t, n\Delta t)$ in the triangle and since an iterative process is needed to resolve the nonlinearity of (6), Lu is replaced by

$$\frac{u_{m,n+1}^{(k+1)} - u_{m,n}^{(k_n)}}{\Delta t} - \frac{1}{2(\Delta x)^2}\Big[u_{m-1,k+1}^{(k+1)} - 2u_{m,n+1}^{(k+1)} + u_{m+1,n+1}^{(k+1)} + u_{m-1,n}^{(k_n)} - 2u_{m,n}^{(k_n)} + u_{m+1,n}^{(k_n)}\Big],$$

where k_n denotes the final number of iterations to obtain an acceptable approximation to $u_{m,n}$ at the grid points on the line $t = n\Delta t$. In the iterative formula two possibilities for the function to replace $f(u^{(k)})$ on the right hand sides of (2) and (4) were considered; one was simply $f(u_{m,n}^{(k_n)})$ and the other was the average $(f(u_{m,n+1}^{(k)}) + f(u_{m,n}^{(k_n)}))/2$. However, since it developed early in the experiments that the latter representation was a bit better than the former and since we were primarily concerned with a comparison of Picard's method with that of Newton, we continued with the latter only for the remainder of the work, which is reported here.

The significant observation was that while the Picard procedure and the Newton procedure required about the same amount of work in the case of the triangle $0 \leq t \leq 1 - x$, $0 \leq x \leq 1$, about nine times the number of iterations required for Newton's method were needed to obtain comparable accuracy by the Picard procedure when the region of interest was the larger triangle $0 \leq t \leq 1.5 - x$, $0 \leq x \leq 1.5$.

The criterion for acceptance of an approximation was a commonly used one, viz., that the maximum relative change per line be less than a prescribed amount before passing to the next line. Thus our requirement was that $k_n = \min k$ such that

$$\max_m \left| \frac{u^{(k)}_{m,n} - u^{(k-1)}_{m,n}}{u^{(k)}_{m,n}} \right| \leq 10^{-6}.$$

Since in this problem the true solution is known (7), we could have a criterion based on the true relative error. However, this is impossible when the solution is not known and we wished to simulate such a condition. A check was made on what the true relative errors were. In the case where the base of the triangle was 1.0 the maximum true relative errors on each line using Newton's method and using Picard's were in agreement with each other to about two significant figures. These maximum true relative errors never exceeded 3.9×10^{-5} and usually lay near the line $t = 1 - 3x/2$. When the triangle had a base equal to 1.5 the solution using Newton's method was run only up to the line $t = 0.23$ and with Picard's up to $t = 0.5$. Again for the values obtained there was very close agreement between the location and the value of the true maximum relative errors. They were generally located near the line $x + t = 1.4$ and generally did not exceed 1.8×10^{-3}. The substantial difference between true relative errors and the relative changes between successive iterations should be taken as a caution to numerical analysts to set bounds in stopping criteria based on relative changes between successive iterations much lower than the desired maximum true relative errors.

The following table gives the comparative numbers of iterations to achieve our criterion for acceptance of an approximation before going on to the next t-line.

BASE OF TRIANGLE	METHOD	NUMBER OF ITERATIONS	t-INTERVAL
1.0	Newton	2 1	(0.01,0.98) (0.98,0.99)
1.0	Picard	4 3 2	(0.01,0.68) (0.68,0.93) (0.93,0.99)
1.5	Newton	3	(0.01,0.23)
1.5	Picard	28	(0.01,0.50)

TABLE 1

3. THE ELLIPTIC CASE

The experiments on the elliptic case were carried out on the numerical solutions of

$$(8) \qquad Lu = u_{xx} + u_{yy} = e^u$$

in the region $0 \leq x \leq 1/2$, $0 \leq y \leq 1/4$, for two sets of boundary conditions $u \equiv 0$ in one case and $u \equiv 10$ in the other. The equation (8) is of considerable interest in some physical problems and the existence and uniqueness of the solutions to the boundary-value problems we have here are assured by the classical theory.

In each of the two cases we compared the two methods for linearization, Picard's and Newton's, which resulted in comparing the numerical solutions obtained for

$$(9) \qquad u_{xx}^{(k+1)} + u_{yy}^{(k+1)} = e^{u^{(k)}}$$

and

$$(10) \qquad u_{xx}^{(k+1)} + u_{yy}^{(k+1)} - e^{u^{(k)}} u^{(k+1)} = e^{u^{(k)}}(1 - u^{(k)}),$$

respectively. As is customary in discretizing the problem for a numerical solution, the Laplacian, $u_{xx} + u_{yy}$, in (9) and (10), was replaced at interior meshpoints of the region of interest by the expression

$$(11) \qquad (u_{m+1,n} + u_{m,n+1} + u_{m-1,n} + u_{m,n-1} - 4u_{m,n})/h^2$$

where $u_{m,n}$ is the discrete analog of $u(m\Delta x, n\Delta t)$ and

$h = \Delta x = \Delta y$, which in the experiments was set equal to 1/64. If we order the points (m,n) such that (m,n) precedes (m',n') if $n < n'$ or if $n = n'$ and $m < m'$ and denote the resulting set of $15 \cdot 31$ numbers $u_{m,n}^{(k)}$ which are the approximations to the solution at $(m\Delta x, n\Delta y)$ at the k-th Picard or Newton iteration by the same notation as above, $u^{(k)}$, then, after division of the component equations by the appropriate factors, viz., the negative of the coefficients of the central term, $u_{m,n}$, in the equations, we obtain systems of linear algebraic equations

$$(12) \qquad (I - L_k - U_k)u^{(k+1)} = b^{(k)}$$

to solve. In (12), I, L_k, and U_k are 465 x 465 matrices, I being the identity, while L_k and U_k are respectively lower and upper triangular matrices appropriate to the particular method of iteration, Picard or Newton. Thus, for our problem, they are transposes of each other and have only two diagonal lines of nonzero elements; in the Picard iteration these elements are identically equal to 1/4, while in the Newton procedure they are equal to $(4 + h^2 \exp u^{(k)})^{-1}$ where the approximation $u^{(k)}$ is evaluated at the central point (m,n) of the star of points indicated in (11). The components of the vector $b^{(k)}$ in (12) are equal to $-h^2/4$ times the appropriate values of the right hand sides of (9) and (10) respectively.

Since in many realistic problems the size of the problems is so huge as to preclude the use of Gaussian elimination to solve the algebraic systems (12), we chose the successive overrelaxation method of Young [11], primarily because it is the simplest iterative method which is substantially better than the Gauss-Seidel method although it is recognized that faster converging block relaxation and alternating direction methods are not much more complicated than successive overrelaxation.

Applying the successive overrelaxation method directly to the systems (12) for a fixed value of k would yield the iterative formula

$$(13) \qquad (I - \omega L_k)u_{r+1}^{(k+1)} = [\omega U_k - (\omega - 1)I]u_r^{(k+1)} + \omega b^{(k)},$$

$$r = 0,1,2,\ldots,$$

where ω is the overrelaxation parameter and the component equations are solved successively in the order which produces the components of the vector $u_{r+1}^{(k+1)}$ consecutively in the order indicated above. However, it is clear that not much effort should be expended in obtaining a very accurate solution to (12) if $u^{(k+1)}$ is not a good approximation to the solution of (8). Similarly, there is not much point to iterating on the index k if the approximations given by (13) are too rough as a consequence of terminating the iteration on r too soon. Thus there exists the open problem as to when one should iterate on k or on r at each step. We avoided the attempt at this relatively difficult analysis and simply iterated on each simultaneously using the formula

$$(14) \qquad (I - \omega L_k)u^{(k+1)} = [\omega U_k - (\omega - 1)I]u^{(k)} + \omega b^{(k)},$$

the component equations being solved in the same order indicated above. Thus we see a strong interweave between the numerical method used for solving either (2) or (4) and the rate of convergence of the actual solutions of (2) or of (4), the difficulty of whose analysis dictates experimentation. We also note that this blend of iterative procedures for the solution of the linear systems (12) and the Newton or Picard iterations to solve the original nonlinear problem has the advantage that one does not have to go to the full-scale effort of solving accuretely the system (12) for each value of k.

As in the parabolic case the criterion for stopping the iterations was that the maximum of the absolute value of the relative change between consecutive iterations of the functional values be no greater than 10^{-6}. If this were truly a bound on the relative error of the numerical solution and the solution of nonlinear equation (8) with the Laplacian replaced by the expression (11[, then this criterion would be about two orders of magnitude too stringent to be justified by the truncation errors. However, some measure of stringency is dictated by the heuristic character of the stopping criterion when one knows neither the true solution nor effective bounds on it.

In the case of Picard iteration where L_k and U_k are independent of k, if one were to ignore the fact that $b^{(k)}$

depends on k the theoretically optimal value of ω is 1.732 for the fastest convergence. For the Newton iteration the value would be a little smaller, again ignoring the change in L_k and in U_k as well as in $b^{(k)}$ from iteration to iteration, which we clearly cannot do. However, in the case of zero boundary conditions for the Picard iteration, where L_k and U_k are constant, the best value (or values) of ω are undoubtedly near the theoretical value noted above since the change in $u^{(k)}$ over the region is very small (The value at the center of the region obtained for the final solution is 0.007071 to four significant figures.). Consequently, for this case we experimented only with the values $\omega = 1.70$ and 1.74. No such confidence existed for the case of Picard iteration on the situation where u is equal to 10 on the boundary because of the large variations in $u^{(k)}$ over the region which would undoubtedly introduce large changes at a point from iteration to iteration. Of course, in the Newton iterations, since L_k and U_k change from iteration to iteration, one has even less of an idea as to the optimal value of ω. Consequently, in the remaining cases a number of runs for different values of ω were made to estimate an optimal value of ω. Table 2 below summarizes the results.

BOUNDARY CONDITIONS	METHOD	VALUE OF ω	NUMBER OF ITERATIONS	VALUE OF $u^{(k)}$ AT (1/4, 1/8)
u = 0	Picard	1.74	64	–0.007071
"	"	1.70	77	"
"	Newton	1.95	258	"
"	"	1.74	64	"
"	"	1.70	77	"
"	"	1.50	154	"
"	"	1.00	426	"
u = 10	Picard	1.55	63	5.65995
"	"	1.50	53	"
"	"	1.40	66	5.65994
"	"	1.35	71	5.65993
"	"	1.30	76	"
"	"	1.20	78	5.65992
"	"	1.10	106	5.65997
"	"	1.00	134	5.65998
"	Newton	1.74	58	5.65995
"	"	1.70	51	"
"	"	1.65	46	"
"	"	1.60	42	"
"	"	1.55	41	"
"	"	1.50	50	"
"	"	1.40	66	5.65996
"	"	1.00	148	5.65998

TABLE 2

From the table we observe that for the case of zero boundary conditions no advantage was obtained from the use of the quadratically converging Newton procedure over the results obtained from the first order converging Picard procedure. In fact, the same values of ω gave the rapidest convergence in both cases. On the other hand, when the boundary conditions were raised to 10, implying rapid changes in u near the boundary, the Newton procedure was more advantageous, requiring some 41 iterations as compared to 53 for Picard's. We note further that curiously the optimal value of ω was slightly lower for Picard's procedure.

As an additional note we included the case of Gauss-Seidel relaxation given by $\omega = 1$ for comparison with successive overrelaxation. As expected, the successive overrelaxation method ranged from two and one-half to almost seven times faster than Gauss-Seidel relaxation.

As an indication of the order of time for a computation, it was observed that on the RAND Johnniac, a Princeton-type machine on which the computations were carried out, it took about eleven seconds for a Newtonian iteration.

Another incidental observation was that the points at which the maximum relative change in $u^{(k)}$ took place were invariably very close to the origin. In most cases it was at the point (1/64, 1/64).

4. CONCLUSION

We have observed in these examples that when the solution of a nonlinear problem has no steep gradients there seems to be no particular advantage in the use of Newtonian methods over those of Picard. On the other hand, when steep gradients occur then there is some advantage, greater in our experiments for the parabolic case than for the elliptic case, where there is considerable interaction between the numerical method of solution of the linearized problem and the convergence rate of the iterated solutions of the linearized problem.

We remark finally that it is possible for the operator L to be quasilinear as well, in which case there are obvious modifications to linearize the nonlinear part of the operator.

We thank Alfred B. Nelson who programmed the experiments for the Johnniac and Ardis McCarroll who collocated the results.

REFERENCES

1. M. R. Hestenes, Numerical Methods of Obtaining Solutions of Fixed End Point Problems in the Calculus of Variations, The RAND Corporation, Research Memorandum RM-102, 14 August 1949.

2. M. L. Stein, On Methods for Obtaining Solutions of Fixed End-Point Problems in the Calculus of Variations, Jour. Res. Natl. Bur. Stand., vol. 50, no. 5, May 1953, pp. 277-297.

3. W. E. Milne, Numerical Solution of Differential Equations, J. Wiley and Sons, New York, 1953.

4. S. A. Chaplyagin, A New Method for Approximate Integration of Differential Equations, recently printed under Classics of the Natural Sciences, Moscow, 1950.

5. L. V. Kantorovich, Functional Analysis and Applied Mathematics, Uspekhi Matematicheskikh Nauk, vol. 3, 1948, pp. 89-185. Also U. S. Natl. Bur. Stand. translation by C. D. Benster.

6. ——————, Certain Further Applications of Newton's Method, Vestnik Leningradskogo Universiteta, vol. 7, no. 2, 1957, pp. 68-103.

7. J. Schröder, Uber das Newtonsche Verfahren, Archive for Rational Mechanics and Analysis, vol. 1, no. 2, 1957, pp. 154-180.

8. R. Kalaba, On Nonlinear Differential Equations, the Maximum Operation, and Monotone Convergence, Jour. Math. and Mech., vol. 8, no. 4, 1959, pp. 519-574.

9. J. Crank and P. Nicolson, A Practical Method for Numerical Evaluation of Solutions of Partial Differential Equations of the Heat-Conduction Type, Proc. Cambridge Phil. Soc., vol. 43, 1947, pp. 50-67.

10. M. L. Juncosa and D. M. Young, On the Crank-Nicolson Procedure for Solving Parabolic Partial Differential Equations, Proc. Cambridge Phil. Soc., vol. 53, 1957, pp. 448-461.

11. D. Young, Iterative Methods for Solving Partial Difference Equations of Elliptic Type, Trans. Amer. Math. Soc., vol. 76, 1954, pp. 92-111.

12. R. Bellman, Functional Equations in the Theory of Dynamic Programming—V: Positivity and Quasilinearity, Proc. Nat. Acad. Sci. USA, vol. 41, 1955, pp. 743-746.